Procedia Computer Science

Volume 66, 2015, Pages 307–316

YSC 2015. 4th International Young Scientists Conference on Computational Science

A quantitative method of text emotiveness evaluation on base of the psycholinguistic markers founded on morphological features

A. Sboev[1,2], D. Gudovskikh[1] R. Rybka[1], and I. Moloshnikov[1]
[1]NRC "Kurchatov Institute", Moscow
[2]MEPhI National Research Nuclear University, Moscow
sag111@mail.ru, dvgudovskikh@gmail.com, rybkarb@gmail.com, ivan-rus@yandex.ru

Abstract
A new quantitative approach to identifying emotionally colored texts that reflect the excited state of its authors is proposed. This approach uses special psycholinguistic markers of text based on morphological characters ratios of Russian language. To apply such markers the morphological parser in combination with an ensemble of SVM classifiers was developed. Testing results of selected topics texts are presented. The study was conducted using texts from different sources such as news, regular blogs, microblogs, social network posts. It showed that results of developed approach can be a useful extension of Big Data methods of traditional sentiment analysis and it can be applicable to developing methods of personality computing.

Keywords: text mining, psycholinguistic markers, text emotiveness, personality computing

1 Introduction

With the increasing popularity and availability of social networks in the world one of the actual problems is the analysis of a large amount of text data. Among the most popular tasks there are extraction and definition of opinion expressed by the author. Opinion mining and sentiment analysis systems (Lee, 2008), (Mohammad Sadegh Hajmohammadi, June 2012) are applicable in various fields of human activities such as advertising, marketing, sociology, political science, forensic linguistic expertise (Olsson, 2008), etc. All of them classify the context of the message. This allows us to understand what the author wrote, but does not allow assessing psychological (emotional) state of the author of the text. In this article an approach is proposed, which aim is to determine the emotional excitation of the author of a Russian text using pre-calculated ratios of morphological characters for the Russian language. The approach allows to determine emotive texts that reflect the excited state of their author. This opportunity would allow to build analytical systems analyzing data from social

Selection and peer-review under responsibility of the Scientific Programme Committee of YSC 2015

doi:10.1016/j.procs.2015.11.036

networks and to solve important problems, for example, assessing the emotional states of society during social and economic disturbances (crises). Such tools, on the one hand, would help to improve the accuracy of the opinion mining systems, and, on the other hand, to carry out research in psycho-social field, using data from social networks and blogs. The task of assessing the state of the author of a text is actively studied by experts in the field of psycholinguistics. The article (Dmitry N. Chernov, 2012) demonstrates correlation between the quantitative characteristics of texts and emotional state of its author. Researchers have attracted a group of students as examinees. The emotional state of examinees was assessed using psychological tests. Quantitative characteristics include a number of morphological properties of words and syntactic features of sentences. A set of quantitative characteristics was formed as a result of processing texts written in different emotional states of examinees. In this work we used the experience of both international researchers, such as Charles Osgood, N. Chomsky, (Doris Aaronson, 2013), and Russian founders of psycholinguistics (A. A. Leontyev, 1997) who have investigated in detail issues of language formation with both psychological features of author and the features of Russian language. This article presents a context-independent approach to determine the emotional reaction on the events discussed in a large set of texts using psycholinguistic markers and the algorithm of morphological ambiguity removal based on Mystem (https://tech.yandex.ru/mystem), working for the Russian language.

2 Materials and methods

The approach presented in this work is based on the psycholinguistic diagnosis of stylistic features of a text with application of methods of natural language processing (NLP). Internet sources provide the following types of markers for texts analysis: markers of activity, lexical markers and another type that we are introducing – psycholinguistic markers.

Activity markers. They reflect the general user activity on the network that is characterized by the number of posts or comments per day.

Lexical markers. There are dictionaries of words, phrases and emoticons in texts that express the authors' opinion. Usually specific dictionaries are prepared in advance for different types of text by experts in a particular domain. Usage of only this type of markers introduces difficulties to analyze texts because same words can have different meanings in the context of various thematic groups of texts.

Psycholinguistic markers. There are indicators that reflect the psychological state of the author at the time of writing the text. Examples of psycholinguistic markers

- The number of pronouns or nouns, adverbs, adjectives or verbs;
- The number of nouns and verbs in comparison with adjectives and adverbs;
- The ratio of the number of verbs to the number of adjectives;
- The number of words;
- The average size of sentences in words;
- The ratio of verbs to the number of nouns;
- The number of exclamation marks;
- The presence of emoticons.

A set of texts with different emotiveness was selected to test this set of markers. Results demonstrated that texts containing nationalist statements with threats had got the marker values that exceeded in several times the values defined by psychologists (see below) while news and users' unemotional comments had normal values of the markers. On the basis of this studies and the results

of correlation analysis (see at the chapter experiments), a core set of three most correlated markers was formed:

- The Trager Coefficient (CT). This is the ratio of the number of verbs to the number of adjectives in the text document. CT is associated with a level of human emotional stability and its normal value is close to 1;
- Coefficient of readiness to action (CRA). This is the ratio of the number of verbs to the number of nouns. This value demonstrates level of human socialization. CRA is interpreted similarly with CT and has the same emotional threshold;
- Coefficient of Aggressiveness (CA). This is the ratio of the number of verbs and their forms (participle and gerund) to the number of all words. Psychologists found that if the value is greater than 0.6 then the author is excited and ready for immediate action.

All selected markers have the common features in the psycholinguistic diagnosis: increased values demonstrate the presence of emotional unrest. It is typical for individuals who are prone to actions; low values indicate such personal characteristics as uncertainty, anxiety.

Sources of information are divided in four main types: news, blogs, social network and microblog. Messages published in the types of data sources have their own characteristics as average length and style of the text. Text size is important in diagnosis, it should not be less than 150 words. For the use of presented markers to the analysis of texts of various types we introduce a complex indicator on the basis of core group of markers (CT, CA, CRA). A more detailed analysis is presented in the Experiments and results part. The indicator is based on the normalized values of the markers (normalization formula) subject to the coefficients of significance. The value of this indicator, which is hereinafter called Total rank, is calculated according to the formula:

$$S = \sum_{i=1}^{n} x_i \cdot a_i \qquad (1)$$

where x_i is the normalized value of the marker i from the core group and a_i is its weight coefficient.

When calculating the Total Rank, Trager coefficient and Aggressiveness are both assigned weights equal to 2. Further the value of the total rank of a message is used for evaluation of its emotiveness. We have calculated the values of the Total rank for some sample texts and have estimated boundaries of emotiveness existence. As it will be shown in Section “Experiments and results”, messages with total rank value above 5 are highly emotive. This range consists of emotional views of the authors expressed on hot themes: geopolitics, sports, relationships.

The samples in the range from 3.5 to 5 consist of texts with less emotiveness, mainly texts from news. Every single text in the range up to 3.5 is advertising, news feeds, digest and informational garbage that does not contain opinions. Later in the work, we evaluate the applicability of the Total rank as a measure of emotive texts.

The process of determining emotively colored text is the following:

- Clearing from noise including html-tags and links;
- Splitting text on tokens (paragraphs, sentences, words);
- Morphological processing by advanced Mystem;
- Shaping the outcomes by computing the values of markers.

The correct definition of the parts of speech is an important part of the estimation of emotively colored text. For morphological parsing of Russian text open source solutions like an AOT, Mystem, etc. are most often used. We chose Mystem for the implementation of the prototype due to its ease of use in python. The accuracy of the parser was evaluated on the texts of the Russian National Corpus (RNC) (http://www.ruscorpora.ru/en/index.html). The results of the experiments showed 47% accuracy of initial algorithm due to morphological disambiguate, hence it needed improvement for high quality. As a result of work Mystem puts all of possible morphological features of words in a list, including weights based on frequency. We define as tag a full set of morphological features of the

word. Evaluation of the completeness of all possible tags from Mystem demonstrated that the system gives the correct options for maximum 94% of the words from the RNC. We have implemented an additional classifier for tags based on support vector machines to achieve this percentage and reduce morphological ambiguity. All sentences are represented as a sequence of words {x1...x2}, where each word is a vector with some characteristics (see below). Punctuation marks are also counted as separate words and are replaced by common tag PUNC. Sequential processing algorithm involves processing proposals from right to left, i. e. from the end of sentence. On each i-th step it takes into description all the known features including the words already parsed on previous steps. A vector is generated for each word. The vector includes the features of the nearest neighbors in the window with size W that is chosen empirically. In this work we use eight words W=(*i*-3, *i*-2, *i*-1, *i*, *i*+1, *i*+2, *i*+3, *i*+4), where *i* is the word that analyzed on step *i*. The feature vector includes the following information for each word:

- All word forms from the window W;
- Tags for those words of W that have been analyzed on previous steps;
- Classes of ambiguities for all words from W (+ their bigrams and trigrams). Class of ambiguity is the set of all possible tags for a word. We represent it as a string of concatenation of the tags. For example, for the Russian equivalent of the word "These" class ambiguity looks like this:
  adjective|nominative_case|plural_adjective|accusative_case|plural|inanimate;
- Possible tags for each word;
- Subtags, i.e. individual morphological features for those words of W that the tags are fixed on the previous steps;
- Possible subtags for each word from the W.

The formation of a full set of morphological features of the words was based on all possible variants from Mystem. We allocated 43 morphological features in accordance with the RNC. We have trained SVM classification model for each morphological feature. The SVM model can give as a result a particular class or the actual number - the so-called decision value, which gives a single score per sample in the binary case. Thus for each word we get 43 positive and negative decision values respectively. We evaluate all possible tags for each word as in formula (2), and the tag that has the highest value *T* will be the winner.

$$T = \sum_{i=1}^{N_i} dv(y_i) + \sum_{j=1}^{N_j} dv(y_j) \quad (2)$$

Here N denotes the number of subtags, *y* – subtag or morphological feature, *dv* – classification decision value, *i* iterates through the number of $y \in x$, *j* iterates through the number of $y \notin x$, where *x* is the set of morphological features included in the tag *x*.

3 Experiments and results

3.1 Testing the morphological tagger

The morphological parsing module was tested on the texts of the National Russian Corpus. The whole corpus at the moment contained 366,245 words. 90% of randomly selected words were used as training set, and the remaining 10% = 38,959 words were used as test set. Table 1 shows the results of the test identification of parts of speech.

Feature name	Number in corpus	Number in testing set	Recall
ADV	23,625	2,363	0.96
CONJ	23,450	2,366	0.95
NUM	3,118	299	0.98
PART	17,556	1,707	0.92
PR	39,007	3,890	1.00
S	158,257	15,771	0.99
V	61,879	6,343	0.99
A	61,372	6,107	0.96
INTJ	59	4	0.50
COM	207	23	0.17

Table 1: Parts of speech classification results

Obviously, some classes (COM, INTJ) were poorly represented in the training sample, and hence low result precision was observed for these ones. Testing the algorithm of full morphological analysis showed accuracy equal to 93.93%. The achieved accuracy coincides with the previously obtained test value for full set of the words in the corpus and meets our needs.

3.2 Experiments with psycholinguistic markers

Unfortunately, the lack of publicly available labeled corpus of emotively colored texts makes impossible a direct comparison of our study with similar ones. Here we present the experiments with markers on samples of different thematic messages collected from 4 types of sources. First of all we performed correlation analysis of marker values. The following set of markers was analyzed:

- The ratio of sum of nouns and verbs to the sum of adjectives and adverbs (mark3p5);
- The ratio of verbs to adjectives (mark_trig);
- The average size of sentences in words (avgsentlen);
- The ratio of verbs to nouns (mark_kod);
- The ratio of prepositions to the total number of words (connection);
- The ratio of prepositions to the total number of sentences

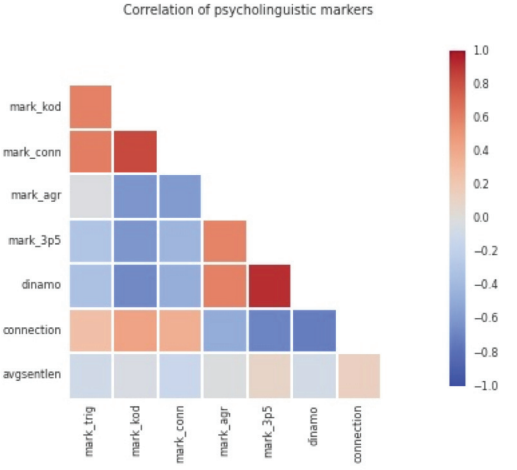

Figure 1: Correlation of psycholinguistic markers.

(mark_conn);

- The ratio of nouns and adjectives to the number of verbs and participles (dinamo).

In the figure 1 positive correlation can be seen between such pairs of markers as CA (mark_agr), CT(mark_trig), CRA(mark_kod). We have grouped them, and further we will call this group Main group. The main group has common psycholinguistic features that are described in the Materials and methods part. There is a negative correlation between all markers of the main group and the marker of dynamism (dinamo), which correlated with coefficient of quality (mark_3p5). A detailed analysis of the texts with the limit values of markers showed that participants of main group could better reflect the emotiveness of texts than other. Therefore further we are focused on the analysis of the main group.

It is logical to assume that the assessment of emotional tension can be regarded as deviation of temporary marker values for a flow of messages on a particular theme. Sample texts on topic of launches of missile “Bulava” published from September 8 to October 17 2014 were collected for the studies. The resulting collection had the size of 13297 texts. As mentioned earlier, texts from different sources have differences that affect significantly the values of the markers. The values of the markers were analyzed in order to reveal these differences separately for the basic 4 types of sources. Preliminary analysis of the average values of markers per day showed that they differ from the diagnostic thresholds specified in psycholinguistics. In particular, daily value of CT was less than the one that psycholinguists have defined of 0.5. The peaks on graph showed that the average values of the markers were less representative then their standard deviation from the point of view of determining increased emotional tension. The standard deviation of marker values reflects the degree of variability of emotional messages on a given topic and is calculated by the formula 3.

$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2}. \quad (3)$$

where N - number of marker values, x_i - markers value, $\bar{x}$ - average markers value.

The following is an analysis of the values of each marker separately for different types of sources: news, microblogging, blogs and social networks. Standard deviations for CRA, CT and CA are shown in Fig 2, 3, 4, respectively.

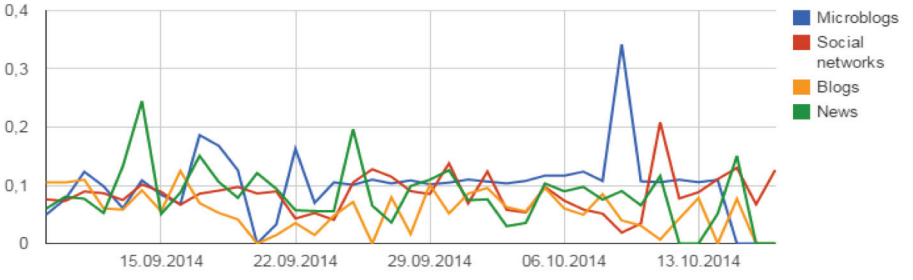

Figure 2: Standard deviation of CRA.

There is a general dependence between the outbreaks in the values of standard deviation of the markers for the entire period, it is shown in figure 4. It is worth noting a strong change in the standard deviation of the CRA marker for texts from news sources. This phenomenon was not observed in the average value of this marker. Strong deviation values are typical for microblogging. Based on the values of the CRA standard deviation during the observed period, we have identified the normal levels of CRA standard deviation for the selected source types: for microblogs it is 0.1, for social networks and news resources it is 0.08 and for blogs it is 0.05. We assume that the values outside the specified standard levels indicate increased emotiveness of the day. We have manually surveyed texts on the dedicated period to assess the significance of markers:

September 14 - the surge was associated with emotional activity in the discussion of the results of launching the missile "Bulava" and the messages about the missiles re-launch. Emotiveness is evident for all types of sources at this day, but it was mainly reflected in the news.

September 16-17 - the increase in the values was due to the reaction of users to the statement of the Russian Deputy Defense Minister about the exact date of re-launching the missiles .

September 25 – the mention of missile "Bulava" appeared in the Ukrainian news sources with emotionally strong anti-Russian context.

September 24-26 - this surge in social networks is associated with the widely discussed topic of nuclear confrontation between the U.S. and Russia. An increase in the number of messages containing the expressed users opinion was observed at the peak of September 26.

During the analyzed period there is a typical weekly activity. Weekly activity is manifested in the growth of deviation values from the beginning of the week till the weekend as for the periods from 15 to 20 September and from 22 to 27 September. When analyzing bursts in the microblogs on September 17 and 22 and October 9, experts did not reveal the presence of emotively colored texts. This indicates the necessity of adapting the methods of psycholinguistics to the peculiarities of microblog format.

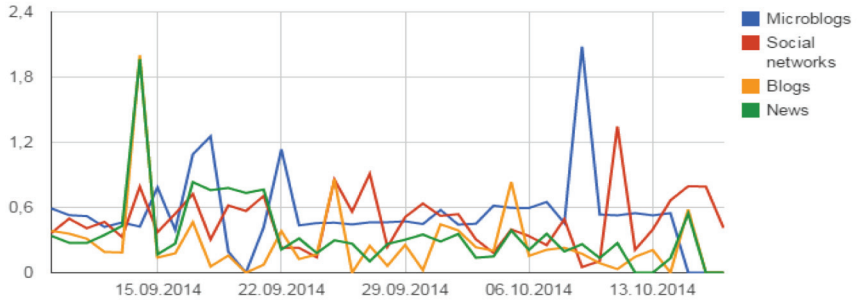

Figure 3: Standard deviation of CT.

As for the Trager coefficient, its values and characteristics coincide with the CRA. It should be noted that for blogs the Trager coefficient more clearly expresses the emotiveness. Based on the values of its standard deviation for the period, we have identified the normal levels of the Trager coefficient standard deviation for the selected source types: for microblogs it is 0.5, for social networks it is 0.45, for news resources it is 0.2 and for blogs it is 0.3. The peaks coincide with the selected CRA, but there are differences for blog posts. The analysis of messages is presented below:

On September 14 and September 25 blogs showed stronger deviations in the Trager coefficient in comparison with CRA. Analysis showed that in these days users were posting quite aggressive messages.

Analysis of minimum on September 20, 26 and 30 showed that among posts in social networks there were no emotional statements and discussions, these days contained mainly news texts. That allows to make a conclusion that the normal standard deviation for Social networks is equal to approximately 0.6.

Peaks in the microblogging for September 18, 22 and October 9 also did not shown any presence of emotively colored texts.

As for aggressiveness, based on the values of its standard deviation during the observed period, we have identified the following normal levels of its standard deviation for the selected source types: 5 for microblogs, 3 for social networks, 3.4 for news resources and 1.5 for blogs.

Analysis of the bursts in the news resources on 13, 14, 21, 25 September and 5 and 9 October did not reveal any presence of emotive texts. Hence the standard deviation of 10 is not a clear indication

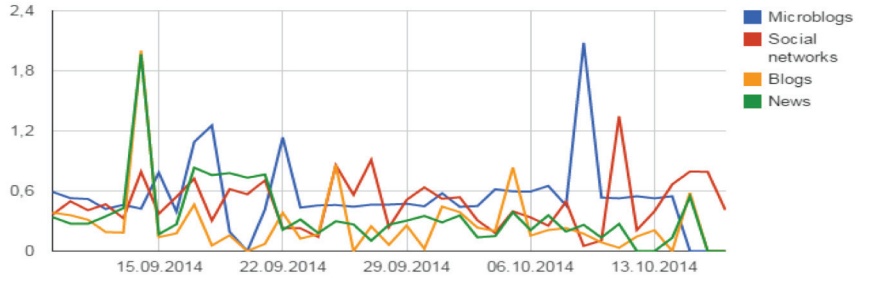

Figure 4: Standard deviation of CA.

of increased emotional background of the day.

September 21, 27 and October 11 – in the texts of social networks exclamation and emotive user reviews were detected. We have concluded that the normal standard deviation for social networks is equal to 4 and that the coefficient of aggressiveness is the most significant for comments in social networks.

October 4 – an interesting proximity in the indicator values for all types of sources was the result of duplicating articles of the same topic about "Russia achieved nuclear parity with the United States for the first time since the Soviet Union decay."

The peaks in the standard deviation of aggressiveness coefficient in microblogging also do not indicate emotively colored texts.

Thus, we concluded that independently the marker of aggressiveness accurately reflects the emotive background only for messages from social networks with daily standard deviation greater than 4.

As a result, following features are highlighted, based on the analysis of daily values of standard deviation of the markers:

- The bursts in the standard deviation values for all markers are more indicative than just averages;
- The proximity of the values of the markers indicates duplicated messages (reposts);
- There is not an explicit relationship between deviations for messages in the microblogging and existence of emotively colored texts in them;
- Average daily values of markers regularly happen to be lower than those specified in the psycholinguistic literature.

- Correlation of each marker for all sources reflects the presence of emotively colored texts in this day.

The results of the evaluation of emotive texts based on the values of each marker individually are not always satisfactory and unequivocal. The obtained features are associated with messages length and text type, because the psycholinguistic methods require the text in the style of an essay or letter, which is not less than 150-200 words. This amount of messages having some opinion is largely provided by the authors on blogs and social networks.

We formed our total rank with given traits of markers of the main group at once. The sensitivity of this indicator can be configured by using weights for different types of sources. In this part of experiments we check how well the total rank shows the emotiveness on the example of texts from blogs and social networks. The texts were chosen with length not less than 150 words.

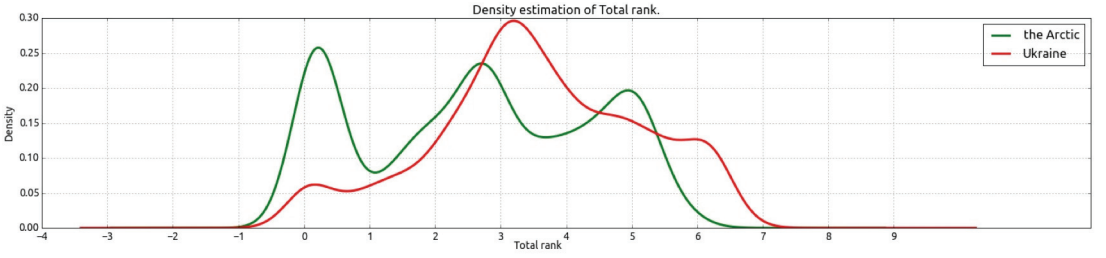

Figure 5: Density estimation of Total rank.

The limitations correspond to the psycholinguistics conditions. The main problem of applying the proposed method lies in the lack of labeled data in public domain. We analyzed values of the selected markers for the various thematic groups, where texts are different in emotiveness. As an example, we chose two relatively polar emotional tensions of the topic "Ukraine" and "Arctic". Nonintersecting sets contain as many messages as 2514 and 4113 respectively. On the figure 5 the graph shows the total rank density estimation for two themes: the Arctic (green) and Ukraine (red).

Analyzing the data sets using this graph, it can be concluded that:

- The theme “Ukraine” contains much more strongly emotive messages with values of total rank more than 4.
- The average values for Ukraine is considerably higher than in the “Arctic” set.
- The theme “Arctic” is not such a quiet theme in the discussions of the Russian Internet community.
- Analysis of the “Ukraine” set of texts with the values of total rank more than 5 showed that it contained letters from soldiers, the situation in the hot spots, furious reviews by Ukrainians and Russians.
- Analysis of the “Arctic” set of texts with the values of total rank more than 5 showed that it contained heated discussions about the development of the Arctic program, articles about the problems of petroleum developments on the shelf.
- Texts with total rank at the range from 3.5 to 5 consist of less emotive reviews and news on both themes.
- The burst on the chart for the topic “Arctic” at values from 0 to 1 is associated with a lot of reposts of an advertising text in which the Arctic was mentioned.

Thus, it can be concluded that the formed complex index «Total rank» does indeed reflect the document emotiveness. Hence, the growing of number of messages with higher values of total rank clearly indicates an increase of the emotional reaction of Internet users to a particular theme.

4 Conclusion

The system to evaluate emotional colors of Russian Internet texts is proposed. This system gives the possibility to investigate the level of social reaction to the selected events. The investigations based on the collections of sample texts for a limited period of time demonstrated that standard deviations of psycholinguistic markers from their intraday values depend on emotive color of texts. The results showed that the selected psycholinguistic markers reflect high emotional reaction of an author, and, in addition, daily values of markers reflect social tensions in relation to an event. A complex indicator reflecting emotiveness of texts on the basis of the core group of markers was presented. On an example of two thematic collections it was shown that on the basis of that complex indicator the most emotional topic could be automatically detected. As shown, proposed approach can serve as an additional analytical tool on the basis of which it is possible to analyze the social-emotive background of topics in the Internet.

5 Acknowledgments

This study was financially supported by the Russian Foundation for Basic Research (project № 15-29-01173 офи_м).

References

A. A. Leontyev. 1997. Osnovi Psicholingvistiki. *Smysl.* 1997, p. 287.

Dmitry N. Chernov, Yuri Y. Ignatov. 2012. Expression of psychological features in speech quantity indicators. *Journal of Psycholinguistics.* 2012.

Doris Aaronson, Robert W. Rieber. 2013. Psycholinguistic Research: Implications and Applications. *Psychology Press.* November 20, 2013.

http://www.ruscorpora.ru/en/index.html. *Russian National Corpus.*

https://tech.yandex.ru/mystem. *Mystem.*

Lee, Bo Pang and Lillian. 2008. *Opinion mining and sentiment analysis.* s.l. : Foundations and Trends of Information Retrieval, 2008. Vol. 2.

Mohammad Sadegh Hajmohammadi, Roliana Ibrahim, Zulaiha Ali Othman. June 2012. *Opinion Mining and Sentiment Analysis: A Survey.* s.l. : International Journal of Computers & Technology, June 2012. Vol. 2.

Olsson, John. 2008. *Forensic Linguistics, Second Edition.* London : Continuum , 2008.